图书情报工作
第64卷 第6期 2020年3月

ChinaXiv合作期刊

英文科技论文摘要的语义特征词典构建

宋东桓[1,2] 李晨英[1] 刘子瑜[1] 韩明杰[1]
[1] 中国农业大学图书馆 北京 100193 [2] 中国科学院文献情报中心 北京 100190

摘 要：[目的/意义]论文摘要是信息组织的重要标引对象，将论文摘要按一定结构进行标引有利于科学传播、知识发现和情报分析。如何对现有非结构式摘要进行精准快速的自动标引是亟待解决的现实问题。[方法/过程]假定不同类别的摘要具有内在一致性，即对结构式摘要的研究可为非结构式摘要自动标引提供方法和技术参考。据此，基于美国国家医学图书馆结构要素标签术语集和标签分类映射关系，提出结构要素 BOMRC 体系和结构式摘要的识别与规范化标引方法。其次选取研究样本并采用文本挖掘方法对样本语料中的单词、动词、三词词块、四词词块等词汇进行词频、TFIDF 值等多个指标的定量统计分析，构建能够进行结构要素识别的语义特征词典。最后利用非结构式摘要测试集进行语义特征词典有效性检验。[结果/结论]结果显示，利用语义特征词典方法能够有效识别非结构式摘要的各类要素，并可用于优化以机器学习方法为核心的自动识别模型。

关键词：科技论文 论文摘要 结构要素 语义特征 特征词典
分类号：G254
DOI：10.13266/j.issn.0252-3116.2020.06.013

chinaXiv:202304.00097v1

1 引言

21 世纪以来，研究论文快速增长，信息过载成为困扰学术界的现实问题。如何让用户快速准确发现所需论文，成为出版界和图书情报界等众多信息服务机构的研究方向。科技论文摘要具有较强的目的性和结构功能，是论文内容的高度概括，也是读者检索和筛选论文的重要依据。同时，论文摘要还是信息组织的重要标引对象，倍受索引数据库重视，对其文本内容的深度挖掘和自动标引也受到图书情报研究和计算机技术应用研究的关注。

目前科技期刊论文摘要存在结构式摘要与非结构式摘要两大类型[1]。相比非结构式摘要存在的格式不固定、层次不够分明、内容不完整、不利于文本挖掘等局限性，结构式摘要在对研究内容表达的完整性、清晰度、信息量、易于移动环境的浅阅读等方面优势凸显，被越来越多的期刊采用。据本课题组 2018 年对 ESI 学科类目下覆盖的 1900 种医学领域期刊，按影响因子排序后采用系统抽样法抽取 20%、即 380 种期刊进行调查发现，有 188 种期刊、占比 49.47% 的期刊采用了结构式摘要。但是采用结构式摘要的科技期刊仍然是少数，对非结构式摘要中出现的研究目的、解决问题的主要方法以及研究获得的重要结果和结论进行分类标引，深入挖掘摘要关键内容仍是信息组织与服务者面临的重要课题。

本研究认为，不同类别的摘要具有内在一致性，即结构式摘要与非结构式摘要在书写体例、惯习用语、写作目标等多方面都有高度一致性。因此，对结构式摘要的研究可以为非结构式摘要自动标引提供方法和技术参考。为此，本研究从现有已采用结构式摘要的期刊论文入手，在总结了 157 种结构式摘要要素标签和 299 种标签组配模式的基础上，提出可映射的“Background-Objective-Method-Result-Conclusion（背景-目的-方法-结果-讨论），简称 BOMRC”要素体系；进而对当前结构式摘要的词汇属性特征进行研究，利用文本挖掘和定量分析构建摘要语义特征词典；最终开发出基于特征词典的摘要标引模型，并在人工标注的非结构式摘要测试语料中进行测试。本研究的价值在于，为结构式摘要的规范化标引和非结构式摘要的结构要素快速识别与标引提供特征词典这一信息组织工

作者简介：宋东桓（ORCID：0000-0001-5671-3796），助理馆员，硕士研究生；李晨英（ORCID：0000-0002-1207-4336），研究馆员，硕士生导师，通讯作者，E-mail：licy@cau.edu.cn；刘子瑜（ORCID：0000-0002-5850-3079），副研究馆员，博士；韩明杰（ORCID：0000-0003-4611-1569），研究馆员，硕士生导师。

收稿日期：2019-07-09 修回日期：2019-09-20 本文起止页码：108-119 本文责任编辑：易飞

具基础,并可用于优化现有自动标引模型和解释自动标引结果,使非结构式摘要自动标引的准确度和可解释性大大上升,为实现千万级别的科技论文摘要标引提供解决方案。

2 相关研究

本研究通过对摘要内容及语义相关研究论文的遍历以及文后参考文献和引文的追踪,共收集到相关英文论文1 526篇,中文论文613篇。研究论文主要发表在计算机、期刊编辑以及应用语言学等领域的学术期刊上,关注的焦点集中在以下两个方面:

2.1 摘要要素的相关研究

通过对重点论文研读并结合关键词共现网络进行分析,可以发现近年来关于科技论文摘要要素的研究主要集中在以下两个方面:

2.1.1 要素的语言特征研究

语言特征大致包括时态、语态、语序、字数以及词汇等,其中时态和语态问题研究侧重于分析摘要写作的新动态,而语序、字数和词汇主要是通过分析一定量的摘要样本获得摘要要素撰写的特征规律。如:曹雁等[2]以"Introduction-Method-Results-Discussions(引言－方法－结果－讨论)、简称IMRD"四要素模式作为分析摘要的对象,利用Range词汇分析软件标记每种要素下词汇,发现每个要素都存在一些带有倾向性的词块。R. A. Day等[3]通过调研各要素中时态的使用频次,发现方法和结果两部分的时态应用较为相似,过去时态使用较为频繁。钱多秀等[4]对论文摘要的各个要素进行对比研究,发现IMRD四要素的时态未来有转向一般现在时的趋势。

2.1.2 要素的模式特征研究

主要侧重于对要素数量和组合的研究,以N. Gratez[5]为代表的学者首先提出了四要素模式,总结出具有普遍性的"Problem-Method-Results-Conclusions(问题－方法－结果－结论)"四要素模式。随后J. M. Swales[6]对N. Graets研究数据获取的可靠性和科学性提出质疑,认为摘要的要素模式应与论文的要素模式一一对应,主张摘要应该由IMRD四要素组成。同时F. Tseng[7]、李涛[8]和周志超[9]等一批学者也都在IMRD模式的基础上,提出了以"Background-Method-Result-Conclusion(背景－方法－结果－结论)、简称BMRC"为代表的其他几种四要素的变体形式。然而一些学者发现为了保证摘要的完整性,应该增加对论文背景的介绍。因此T. Dahl[10]基于J. M. Swales的研究提出了"Background-Purpose-Methodology-Result-Comments on results(背景－目的－方法－结果－结果解释)"五要素模式。此外,医学领域论文早在1987年R. B. Haynes[11]就提出了"Objective-Design-Setting-Patients or participants -Interventions-Measures and Results-Conclusion(目的－设计－地点－患者或参与者－干预－测量和结果－结论)"的七要素模式,目前许多医学期刊根据文章类型给出了多种类型的结构式摘要撰写要求,例如:JAMA Surgery、Physiotherapy要求的结构式摘要中最多有8个要素。

2.2 摘要语义特征的相关研究

美国语义学专家L. F. Don和A. P. NILSEN[12]提出语义特征包括五大类,分别为:语法－语义特征、内在语义特征、谓语性语义特征、状语性语义特征和感受性语义特征。对词汇或者其他实体进行语义特征分析时,往往使用"[+ －语义属性]"来表示对应的语义特征。第2类和第5类属于词汇层面的语义特征,要素类别可以充当语义特征属性,如:[+ background][－objective][－method][－result]和[－conclusion]。其余类属于语法层面的语义特征分析,不能脱离句子而分析,单个词汇不能表现出任何语义。其中利用语义特征技术实现论文摘要要素识别的研究包括:基于单一特征的语义识别技术研究和基于综合特征的语义识别技术研究。

基于单一特征的语义识别技术研究是指仅仅利用词频、语序、时态等某个特征进行摘要要素的语义识别。2002年L. E. ANTHONY[13]首次构建出摘要自动识别模型,他最初是利用少数的摘要数据,从摘要数据中提取一到五个单词的连续单词集群,基于朴素贝叶斯算法进行学习,以达到摘要结构要素内容的识别。而S. N. Kim和L. MARTINEZ[14]通过研究发现应用语序进行结构要素识别时,条件随机场算法要比朴素贝叶斯算法和支持向量机效果更好,精确度一般在90%以上。而综合特征的语义识别技术相比单一特征的语义识别技术对人的主观性依赖较多,需要人为选择待分析的特征。如V. D. Feltrim等[15]将摘要划分为若干句子群,通过对句子所处位置进行结构要素识别研究。J. Silva等[16]和Y. K. Meena等[17]也利用句子特征构建了不同类型的要素识别模型。Y. Guo等[18]利用句法分析工具对词汇特征和语境特征的效果进行比较,发现词汇特征的预测效果最好,语态和要素语序的识别效果最差。沈思等[19]以摘要文本中的字为基本语义单位,基于LSTM-CRF模型的深度学习

chinaXiv:202304.00297v1

ChinaXiv合作期刊

方法构建出期刊论文摘要结构功能自动识别模型。但其结构要素标签的选择并没有考虑到学科的差异性。

综上发现:①词块具有独特性特点,词汇特征的预测效果比语态和要素语序的识别效果更好;②词汇属性可以视为语义特征体现,同时内在语义特征主要侧重于基本概念、基本逻辑的语义特征;③摘要要素识别的研究,重点关注了词频、时态、语态、位置等语义特征。但从词汇属性的角度,考虑构建语义特征词典进而完成摘要内容标记的研究尚未见报道。因此为解决以往研究中主观依赖性强、特征稀疏以及可解释性受限等问题,本研究试图以定量分析为主线、以词汇属性为引导构建出语义特征词典,为摘要要素的识别建立基础。

3 研究方法与数据准备

3.1 研究目标和设计

为了提高信息组织的智能化标引水平,为计算机进行结构式摘要的规范化标引与非结构式摘要的结构化标引以及情报分析中的信息抽取提供参考,本研究以英文科技期刊论文的结构式摘要数据为样本,通过对以下三个问题的具体研究,深入挖掘结构式摘要的结构要素及其文本特征,以构建具有结构要素识别功能的语义特征词典:①如何确定结构式摘要识别与标引方法?②结构要素中是否存在具有语义识别功能的代表性特征词汇?③特征词是否能识别非结构式摘要中句子的结构要素,识别效果如何?本研究首先对结构式摘要特征进行分析总结,发现 35% 以上的结构式摘要采用了 BMRC 或 OMRC 的标签组配模式,同时发现美国国家医学图书馆提供的结构式摘要标签术语集中也对标签按 BOMRC 进行了标签分类关系映射,因此提出了采用 BOMRC 五要素模式进行结构式摘要识别及标引的方法。然后通过对统一映射到 BOMRC 模式的结构式摘要句子的内容特征词计算,完成了不同结构要素下特征词候选集的提取,并采用结构式摘要测试集完成特征词候选集的修正与完善工作,构建了适用于 BOMRC 结构要素标记的语义特征词基础词典,最后利用语义特征词典进行识别有效性检验工作。具体研究内容与研究设计如图 1 所示:

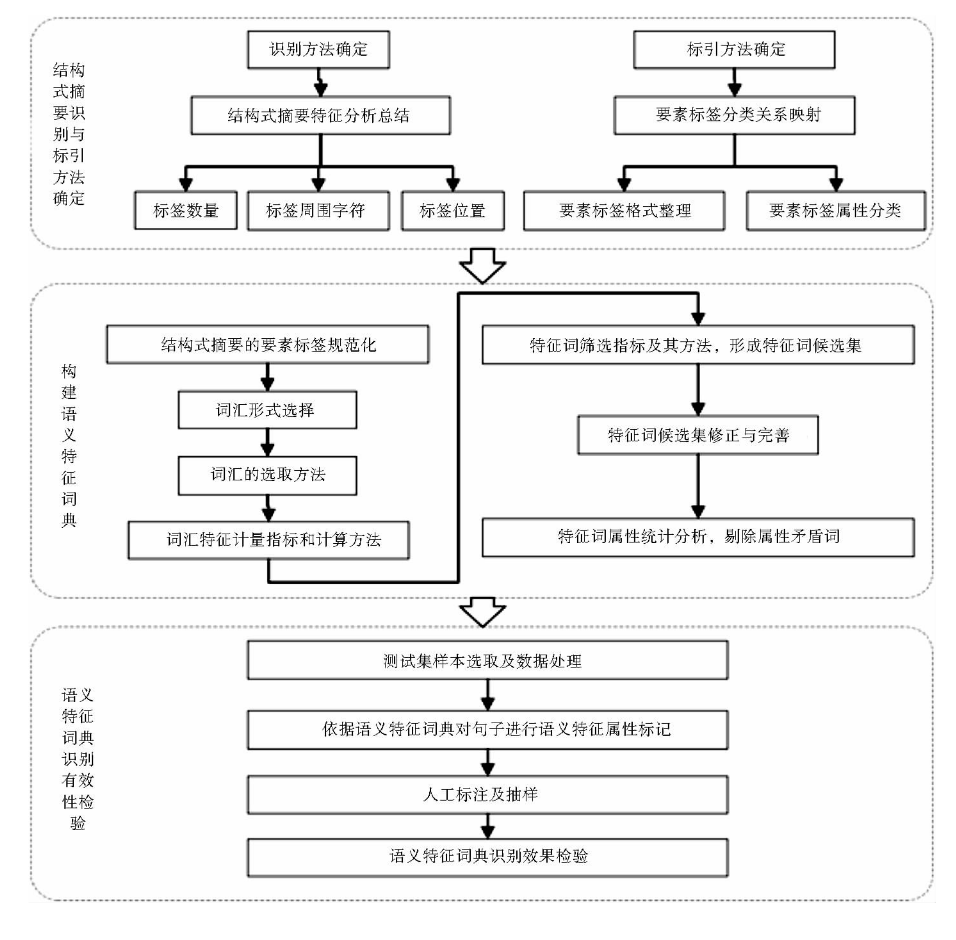


图 1 研究设计思路

chinaXiv:202304.00297v1

3.2 数据准备

3.2.1 期刊论文基础数据准备

本研究采用科睿唯安公司基于 WOS 核心合集数据库,为“全球工程前沿研究项目”标记的中国工程院下属 9 个领域的 TOP 10% 高被引论文及其施引论文数据 284 525 条(2012－2017 年);剔除非 Article 文献类型数据,按期刊抽取每种期刊的最新一期论文数据,实际得到 10 143(去重后为 7 218)种期刊的 16 900(去重后为 13 046)篇论文简要信息数据,如表 1 所示:

表 1 中国工程院 9 个领域下期刊基础论文数据准备过程及其筛选结果

学部	高被引论文及其施引论文篇数	来源期刊种数	最新期刊论文篇数
1 机械与运载工程	26 804	1 033	2 102
2 信息与电子工程	24 716	710	1 586
3 化工、冶金与材料工程	26 047	519	1 103
4 能源与矿业工程	31 065	765	1 584
5 土木、水利与建筑工程	27 135	1 033	1 908
6 环境与轻纺工程	30 942	1 204	1 922
7 农业	24 102	1 438	2 070
8 医药卫生	22 254	1 660	2 414
9 工程管理	18 870	1 681	2 211
合计	231 935	10 043(7 218)	16 900(13 046)

3.2.2 结构式摘要论文筛选

(1)以美国国家医学图书馆提供的 3 032 个结构式摘要标签作为标签术语集[20],结合本研究过程中收集的结构式摘要要素标签后跟随的特殊标记字符特征集,采用简单的模式匹配和前方一致匹配相结合的方法,依据摘要中出现的标签位置和数量等特点进行判断,对13 046论文摘要数据进行筛选,共筛选出1 583篇采用结构式摘要的论文,如图 2 所示:

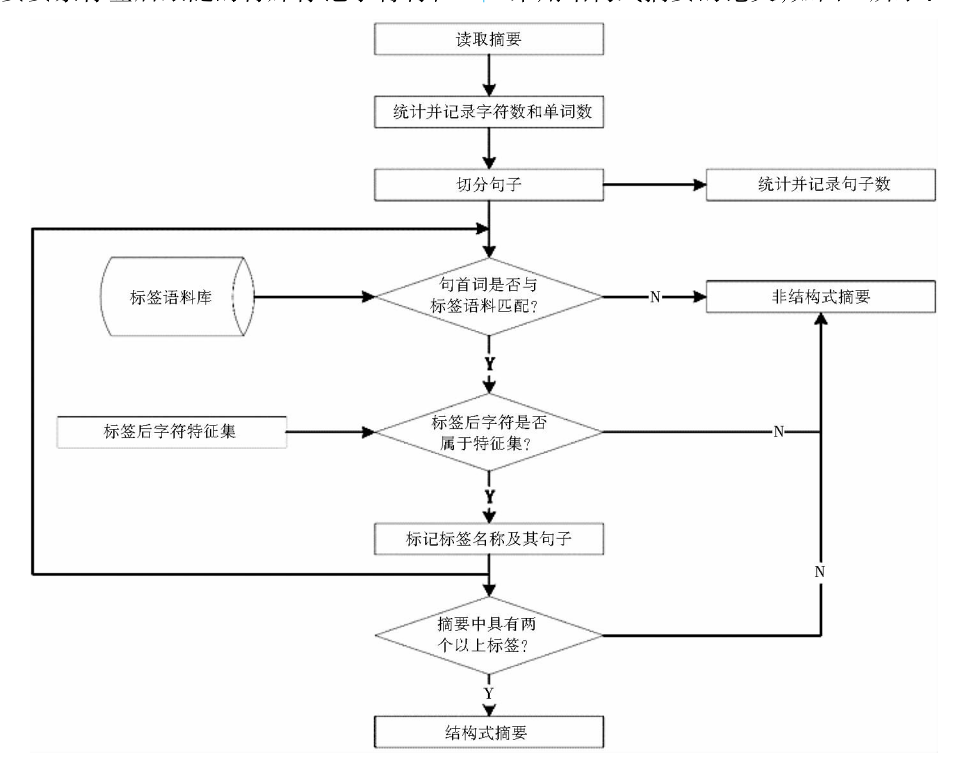


图 2 结构式摘要论文筛选流程

(2)对拟筛除的非结构式摘要进行人工核查发现,个别结构式摘要中存在:标签拼写错误、特殊标签格式、标签未被美国国家医学图书馆提供的标签术语集收录等问题。人工补充 11 篇未准确标记的论文后,共标记出 1 213 种期刊的 1 594 篇较新论文作为结构式摘要研究样本。结构式摘要识别及标引结果样例,如图 3 所示:

chinaXiv:202304.00297v1

ChinaXiv合作期刊

原始数据

```
Background: We aimed to define the depth and time of maximal anti-tumour response to programmed death-1 blockade antibodies
(anti-PD1) in heavily patients with classical Hodgkin lymphoma (HL). To this end, we evaluated the kinetics of response for
up to two years. Materials and methods: The 18F-FDG positron-emission tomography (PET) and contrast-enhanced tomography (CECT)
data of all relapsed or refractory HL treated at France, from 2013 to 2015 were retrospectively reviewed according to the
International Project 2014 criteria and the Response to therapy Criteria (LYRIC) Results: Sixteen patients were included.
The median (range) treatment duration was 18.4 (2.8-23.7) months. Fifty-six percent of patients (9/16) achieved an objective
response at 3 months, including 19% (3/16) of complete response. Seventeen percent (1/6) of partial at 3 months were converted
in a complete response. 22% (2/9) of 3 months relapsed before one year. The nadir was reached at 12.7 (3.0-23.0) months. The median
(range) depth of response at nadir was -77% (-50% to 100%). Conclusion: We concluded that complete metabolic responses occurred
within 6 months, a minority of partial responses were converted in complete response, and the median nadir was observed one year
after treatment initiation. These data could help to better define the optimal treatment strategy by PET or approaches. (C) 2017
Elsevier Ltd. All rights reserved.
```

规范化后的数据

```
<background> We aimed to define the depth and time of maximal anti-tumour response to programmed death-1 blockade antibodies (anti-PD1)
    To this end, we evaluated the kinetics of response for up to two years.
<materials and methods> The 18F-FDG positron-emission tomography (PET) and contrast-enhanced tomography (CECT) data of all relapsed or
<results> Sixteen patients were included. The median (range) treatment duration was 18.4 (2.8-23.7) months.
    Fifty-six percent of patients (9/16) achieved an objective response at 3 months, including 19% (3/16) of complete response.
    Seventeen percent (1/6) of partial at 3 months were converted in a complete response.
    22% (2/9) of 3 months relapsed before one year. The nadir was reached at 12.7 (3.0-23.0) months.
    The median (range) depth of response at nadir was -77% (-50% to 100%).
<conclusion> We concluded that complete metabolic responses occurred within 6 months, a minority of partial responses were converted in
    These data could help to better define the optimal treatment strategy by PET or approaches.
```

图 3　结构式摘要识别及标引结果样例

（3）选取数据集中 1 213 种期刊包含的全部13 781 篇论文数据，再次进行结构式摘要筛选，其中含有非结构式摘要论文 5 021 篇（作为非结构式摘要检测集），结构式摘要论文 8 760 篇。排除结构式摘要论文中覆盖的 1 594 篇前期研究样本数据，其余 7 166 篇作为后续研究的结构式摘要测试集。同时对标签术语集未收录的标签进行补充和映射关系对应，共补充 30 个标签。

4　语义特征属性分类研究

4.1　结构要素的标签数量分布

对 1 594 篇论文的结构式摘要采用的 157 种要素标签进行统计发现，所有标签出现了 6 582 次，平均每个标签出现在 42 篇摘要里。其中 Conclusion 出现的频次最高，接下来依次是 Result、Method、Background 和 Objective，如图 4 所示：

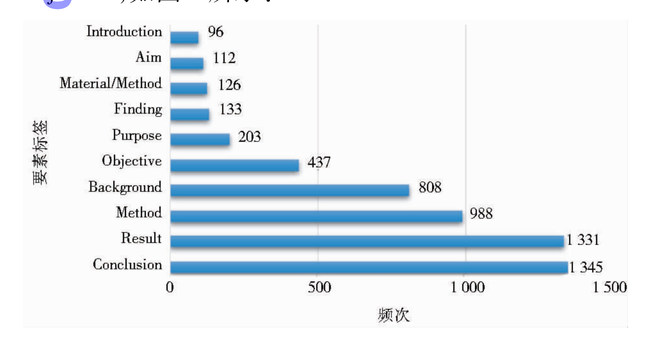


图 4　结构式摘要要素标签出现频次 TOP10

4.2　结构要素的标签组配模式分布

通过对标签组配模式进行统计发现，共出现了 299 种类型。其中，出现频次最多的是“Background + Method + Result + Conclusion、简称 BMRC 模式”，占比超过 1/4；其次是“Objective + Method + Result + Conclusion、简称 OMRC 模式”和“Background + Result + Conclusion、简称 BRC 模式”，如图 5 所示：

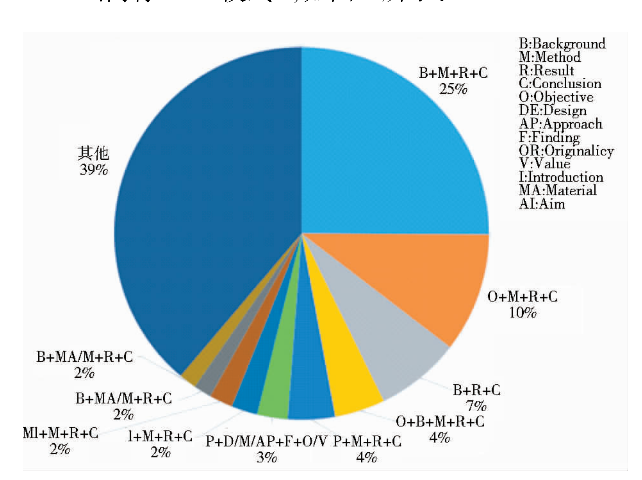


图 5　结构式摘要的结构要素标签组配模式分布

该结果摒弃了对单要素占比情况的分析考察方式[9]，从整体上对组配模式的占比做了详细统计，这样可以更有利于挖掘要素之间的顺序性。同时发现 Conclusion + Result + Method + Background + Objective 五要素之间的组合最为普遍。本研究也通过调研大量期刊投稿说明以及文献中各研究要素的定义，对 BOMRC 五要素包含的概念做了总结（见图 6）。根据规定的要素定义，发现结构式摘要所有的要素标签都可以映射到该五要素下。这不仅可以保证摘要内容识别的完整性，还可以区分出摘要的核心内容，因此利用词汇属性对 BOMRC 五要素内容进行识别具有很重要的意义。

5　语义特征词典的构建

本研究期望获取可对非结构式摘要文本内容进行结构化识别与标引的特征词汇，而非结构化摘要文本内容特征的识别以句子为单位相对简单易行，因此将前期获得的结构式摘要文本全部以句子为单位作为文

chinaXiv:202304.00297v1

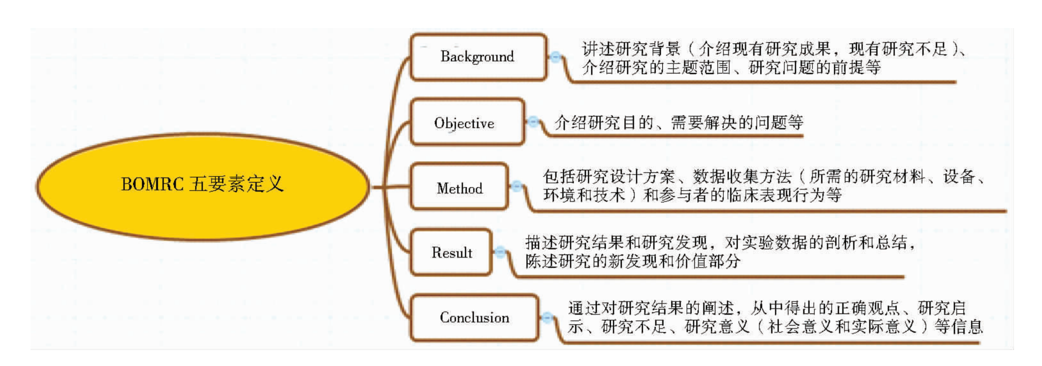


图 6　BOMRC 五要素定义

档，并结合美国国家医学图书馆归纳的映射关系，对提取的要素进行 BOMRC 五要素归类，以期筛选出具有识别句子所属结构要素功能的特征词汇。另外，TFIDF 方法是当前研究中最常用的词文本特征加权方法，它较好地将词的局部权重同全局权重结合在一起，可识别出在一篇文档中出现次数多而在整个文档集合的其他文档中出现次数少的词。因此，本研究通过以下流程和计算方法进行了特征词筛选。

（1）以 BOMRC 五要素分类结果为文档集，即每个结构要素下的句子为一个文档，分别识别和标记句子内的单词（剔除动词）、动词、三词词块和四词词块，排除数字、符号等非英语词汇字符，统计其频次。

（2）根据 BOMRC 每个文档集中文档标记单词、动词、三词词块和四词词块等词汇的频次以及包含这些词汇的文档数、总文档数，计算每个词汇分别在 BOMRC 五要素文档集中的 TFIDF 值，计算公式如图 7 所示：

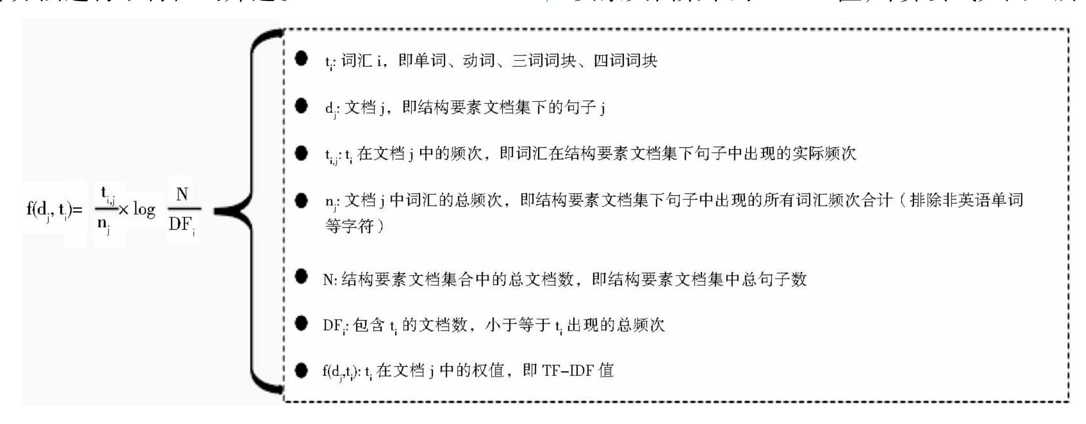


图 7　TFIDF 值计算指标确定

5.1　语义特征词典候选集的构建

根据每个词汇在 BOMRC 五要素文档集合中出现的频次及其 TFIDF 值占比，人工观察词汇各指标的数值及其区间分布状况，发现按照以下阈值筛选出有可能作为识别句子结构化功能的特征词（见表 2）：

- 单词：词频≥5 且 TFIDF 值占比≥50%（排除其他四个要素中存在 TFIDF 值≥40%）
- 动词：词频≥3 且 TFIDF 值占比≥50%（排除其他四个要素中存在 TFIDF 值≥40%）
- 三词词块：词频≥2 且 TFIDF 值占比≥50%（排除其他四个要素中存在 TFIDF 值≥40%）
- 四词词块：词频≥2 且 TFIDF 值占比≥50%（排除其他四个要素中存在 TFIDF 值≥40%）

通过对初筛指标的定义，从 481 877 个词汇中筛选出 15 526 个特征词，形成特征词候选集。但由于存在包含关系的特征词会出现属性不一致的情况，因此需要对单词－三词、单词－四词、动词－三词、动词－四词和三词－四词等五种情况进行分析，剔除属性不一致的特征词（例如："suggest"的属性为 Conclusion，而"recent studies suggest that"的属性为 Background），共剔除 451 个特征词，保留 15 075 个特征词（见表 3）。

5.2　语义特征词典候选集的修正

特征词标注准确率是特征词典识别效果的核心因素，因此如何提高特征词候选集中词汇的平均标注准确率是本节关注的重点问题。以 7 166 篇论文的结构式摘要测试集作为语义特征词典候选集修正与完善的语料，分句后剔除与摘要内容无关的句子内容，如版权信息、链接和邮箱等，共得到 80 346 个句子。对每个句

chinaXiv:202304.00289v1

ChinaXiv合作期刊

表 2 结构式摘要 BOMRC 五要素文本中特征词汇筛选结果

词汇类型	数量和占比/%	Background	Objective	Method	Result	Conclusion
单词	总词数	7 087	5 718	9 336	10 779	8 309
	特征词数	64	23	183	95	47
	占比	1.16	0.70	3.29	2.08	1.22
动词	总词数	1 637	1 394	1 983	2 357	1 971
	特征词数	56	43	130	120	89
	占比	3.60	3.08	6.86	5.77	4.72
三词词块	总词块数	36 778	23 538	49 943	64 788	48 120
	特征词数	1 448	924	2 594	3 393	1 711
	占比	3.94	3.93	5.19	5.24	3.56
四词词块	总词块数	35 349	22 593	45 070	58 843	46 229
	特征词数	643	479	1 071	1 645	768
	占比	1.82	2.12	2.38	2.80	1.66

表 3 剔除属性矛盾词后的特征词候选集的属性分布

词汇类型	Background	Objective	Method	Result	Conclusion	总计
单词	57	20	175	89	44	385
动词	55	38	122	116	84	415
三词词块	1 399	874	2 556	3 321	1 669	9 819
四词词块	594	469	1 043	1 615	735	4 456
总计	2 105	1 401	3 896	5 141	2 532	15 075

子中出现的特征词及相应的标注标签准确性进行统计（正确标记为 1，错误标记为 -1，未标注标记为 0），同时对每个特征词整体标注的准确性占比进行区间划分，利用特征词典候选集对结构式摘要句子内容标记所得结果与结构式摘要句子本身所具有的标签进行核对。将标注准确率小于 50% 的特征词作为剔除对象，共保留 6 447 个特征词。表 4 是 4 种类型特征词汇标记准确率在不同区段的词汇数量，显然三词特征词块的标记准确率明显高于其他类型的特征词汇。

表 4 4 种类型特征词汇标记准确率在不同区段的词汇数量分布

词汇类型	词汇类型	> =90%	80% -90%	70% -80%	60% -70%	50% -60%	<50%	总计
Background	①	0	0	0	1	3	52	56
	②	1	0	1	1	6	45	54
	③	136	16	28	67	137	654	1 038
	④	78	8	7	24	62	193	372
Objective	①	0	1	0	1	1	16	19
	②	0	1	0	0	3	33	37
	③	66	17	20	55	91	425	674
	④	65	9	25	54	69	138	360
Method	①	7	16	34	26	26	63	172
	②	4	9	19	24	24	42	122
	③	684	206	142	165	187	462	1 846
	④	312	70	34	46	34	104	600
Result	①	4	6	13	9	17	34	83
	②	13	15	23	15	15	28	109
	③	721	236	173	212	171	441	1 954
	④	370	87	69	66	48	112	752
Conclusion	①	1	0	2	6	7	26	42
	②	2	1	6	8	12	55	84
	③	277	98	77	115	132	517	1 216
	④	148	36	27	35	51	155	452

注：①特征单词；②特征动词；③三词特征词块；④四词特征词块

chinaXiv:202004.00297v1

宋东桓，李晨英，刘子瑜，等.英文科技论文摘要的语义特征词典构建[J].图书情报工作,2020,64(6):108－119.

5.3 语义特征词典候选集的完善

由于特征词的选取主要依据词频、TFIDF 值的占比两个要素,因此修正后候选集词汇的基本特征可以从特征词频次占比、特征词 TFIDF 值排名区间占比两个方面进行分析。具体分析指标值的计算过程如下:

频次占比:①对每个要素中出现的所有词汇进行频次统计;②计算每个要素下词汇频次在总语料中同一词汇频次的占比;③统计特征词汇的频次占比区间。

TFIDF 值排名区间:①计算每个要素中出现的所有词汇的 TFIDF 值;②对同一要素下所有的词汇按照 TFIDF 值的大小由低到高进行排名并编号;③对所有编号进行归一化,统计 TFIDF 值的排名区间,要素排名区间 = 该要素下词汇的排名值/该要素下词汇总频数。

对 BOMRC 五要素中包含修正后特征词汇频次占比和 TFIDF 值排名区间进行计算,如图 8(其中 1:Background、2:Objective、3:Method、4:Result、5:Conclusion)。发现特征单词、特征动词、三词特征词块和四词特征词块的频次占比主要集中在 65% －100%、60% －100%、60% －100% 和 60% －100%;而 TFIDF 值排名区间在前 30%、前 45%、前 10% 和前 5%,因此可以综合词频、词频占比、TFIDF 值及 TFIDF 值排名区间四个特征指标作为特征词的补充标准。而由于测试集的句子数量增加,需要对词频作出调整,结合人工观察最终确定单词词频≥100,其余类型的词汇词频≥5 作为词频的新标准。利用结构式摘要测试集计算其中出现的词汇及其计量指标,最终补充了 5 542 个特征词,此时词典中共计11 989个词汇。经过属性统计,剔除存在包含关系的特征词属性不一致现象,最终确定了 11 761个特征词,即完善后特征词典的词汇量扩充到 11 761 个。

chinaXiv:202304.00297v1

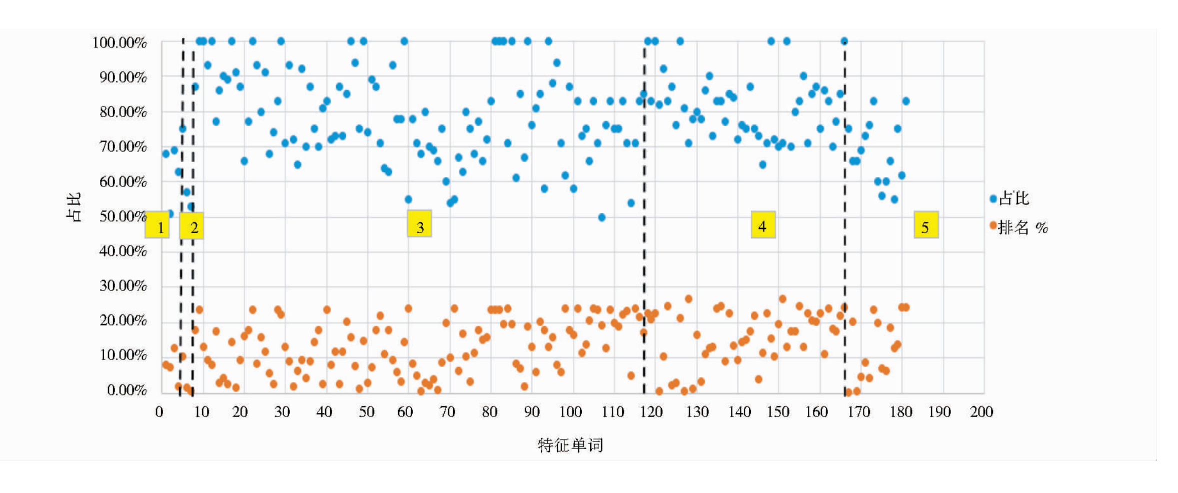


图 8－1　各要素下特征单词频次占比及 TFIDF 值排名区间分布

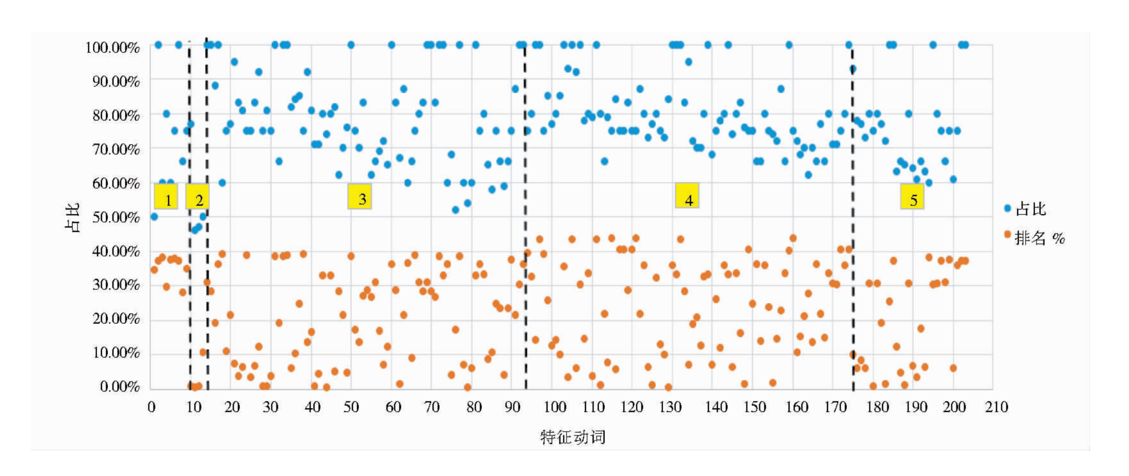


图 8－2　各要素下特征动词频次占比及 TFIDF 值排名区间分布

ChinaXiv合作期刊

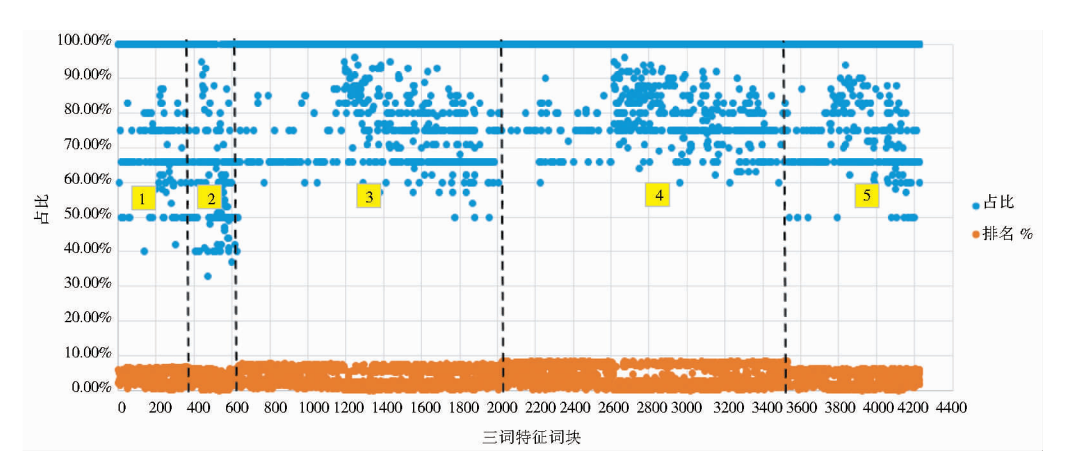


图 8－3　各要素下三词特征词块频次占比及 TFIDF 值排名区间分布

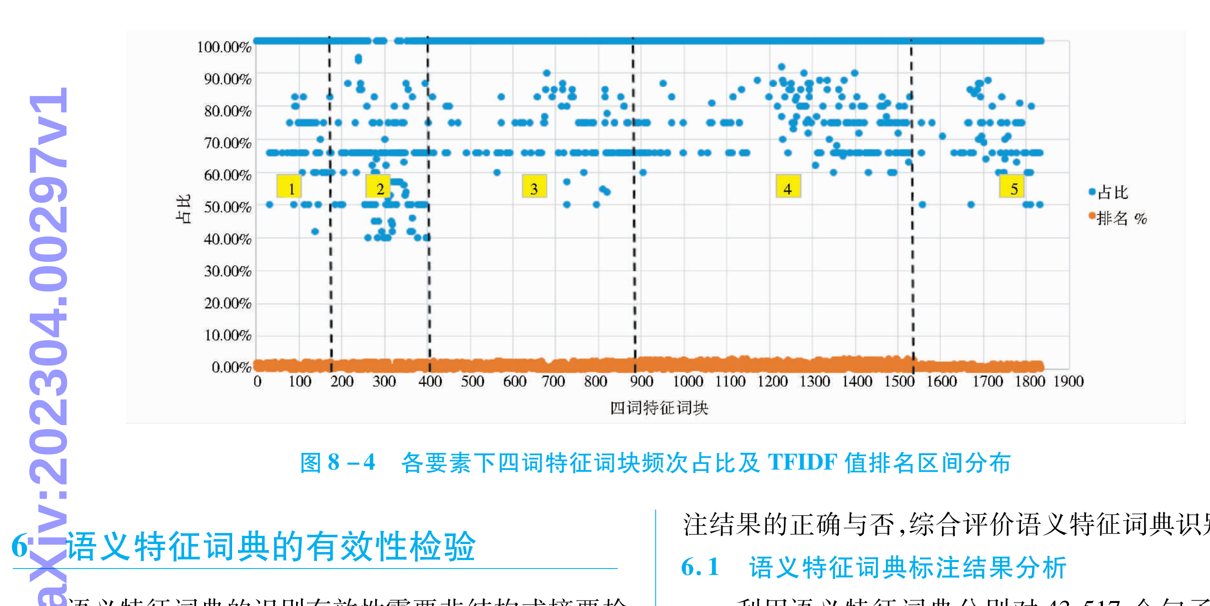


图 8－4　各要素下四词特征词块频次占比及 TFIDF 值排名区间分布

6　语义特征词典的有效性检验

语义特征词典的识别有效性需要非结构式摘要检测集进行验证。因此将已提取的 5 021 篇非结构式摘要数据作为检测集，通过分句操作和剔除不相关句子，共获取 43 517 个句子。使用语义特征词典对每个句子进行要素标注，通过抽样结合人工检验判断句子标注结果的正确与否，综合评价语义特征词典识别效果。

6.1　语义特征词典标注结果分析

利用语义特征词典分别对 43 517 个句子进行精确匹配，打上相应要素标签。据统计，共对 29 530 个句子标注了标签，占比 67.86%。11 761 个特征词参与了机器标注，占比 73.12%。具体匹配的特征词数如表 5 所示：

表 5　BOMRC 五要素下参与机器标注的特征词数量分布

匹配特征词数	Background	Objective	Method	Result	Conclusion	总计
特征单词	7	1	123	55	19	205
特征动词	13	5	90	92	30	230
三词特征词块	723	406	1 633	1 893	1 393	6 048
四词特征词块	271	289	491	617	449	2 117
总计	1 014	701	2 337	2 657	1 891	8 600

另外对标注的句子数进行统计发现（见表 6），利用三词特征词块标注的句子数最多，而四词特征词块标注的句子数最少。而通过对句子被标注标签的观察发现，“Method + Result”的标签组合形式最多（2 552 个）。说明句子中同时出现 Method 与 Result 要素的特征词，经分析发现部分原因在于 Method 中会出现 Outcome、Outcome measurement 等表征临床试验结果的干扰要素。标注结果样例见图 9。

表 6　BOMRC 五要素的特征词标注句子数量分布

匹配特征词数	特征单词	特征动词	三词特征词块	四词特征词块	总计
Background	768	213	2 749	1 251	4 981
Objective	61	547	1 206	794	2 608
Method	4 409	2 619	3 171	846	11 045
Result	2 417	3 621	6 566	2 114	14 718
Conclusion	1 370	2 226	4 841	1 951	10 388
总计	9 025	9 226	18 533	6 956	43 740

ChinaXiv:202304.00297v1

原始数据：
Nonalcoholic fatty liver disease (NAFLD), defined by excessive liver fat deposition related to the metabolic syndrome, is a leading cause of progressive liver disease, for which accurate non-invasive staging systems and effective treatments are still lacking. Evidence has shown that increased ferritin levels are associated with the metabolic insulin resistance syndrome, and higher hepatic iron and fat content. Hyperferritinemia and iron stores have been associated with the severity of liver damage in NAFLD, and iron depletion reduced insulin resistance and liver enzymes. Recently, Kowdley et al demonstrated in a multicenter study in 628 adult patients with NAFLD from the NAFLD-clinical research network database with central re-evaluation of liver histology and iron staining that the increased serum ferritin level is an independent predictor of liver damage in patients with NAFLD, and is useful to identify NAFLD patients at risk of non-alcoholic steatohepatitis and advanced fibrosis. These data indicate that incorporation of serum ferritin level may improve the performance of noninvasive scoring of liver damage in patients with NAFLD, and that iron depletion still represents an attractive therapeutic target to prevent the progression of liver damage in these patients. (c) 2012 Baishideng. All rights reserved.

机器标注的数据：
<background>Nonalcoholic fatty liver disease (NAFLD), defined by excessive liver fat deposition related to the metabolic syndrome, is a leadin
<background>Evidence has shown that increased ferritin levels are associated with the metabolic insulin resistance syndrome, and higher hepati
<background>Hyperferritinemia and iron stores have been associated with the severity of liver damage in NAFLD, and iron depletion reduced insu
<result>Recently, Kowdley et al demonstrated in a multicenter study in 628 adult patients with NAFLD from the NAFLD-clinical research network
<objective;;conclusion>These data indicate that incorporation of serum ferritin level may improve the performance of noninvasive scoring of li

图 9　非结构式摘要标引结果样例

6.2　人工标注及抽样

进行机器标注之后，本研究邀请两位标注者进行结果核对。首先抽取 10 篇摘要，向标注者展示如何进行人工标注。由于仅限于对 BOMRC 五要素进行标注，因此将前面提到的五要素定义作为标注标准。经过培训后，邀请两位标注人员独立对同样的 20 篇非结构摘要进行先验标注，观察两人标注的结果。对标准不一致的结果进行沟通，最终达成一致意见。根据上述标注方法和标准，将下载的 5 021 篇非结构式摘要数据按照所属领域进行划分，对每个领域的数据首先按照论文数的占比进行划分，之后对每个领域中按期刊 ISSN 号进行升序排列，等距抽取对应期刊（组距＝5），分别筛选出 4、4、4、2、4、6、26、40、8 本期刊。对每本期刊中的论文摘要数据按照论文元数据的唯一标识符进行升序排列的第一篇论文摘要作为人工标注的对象。按上述标准共选取 98 篇论文数据，进行人工标注。

6.3　语义特征词典识别效果校验

评估是信息检索、机器学习和自然语言处理领域必要的工作之一，目前常用准确度、精确率、召回率及 F1 值等指标来对模型或词典的综合识别效果进行判断。通过对 4 种特征词汇的语料所标注的标签与论文摘要中的原始标签进行核对，获取每个特征词汇的识别情况（见图 10）。综合 4 种类型特征词汇求取平均值发现，4 种特征词汇对五要素的识别准确率都能保证在 85% 以上，其中对 Objective 的识别准确率最高，达到 90%。在精确率方面，对五要素的识别效果相差不大，精确率都能保持在 80% 以上。而从召回率上看，Background 和 Objective 的召回效果较差，Method 的召回率较为客观。另外 F1 值的平均数大小可以反映对五要素识别效果的综合评价，结果显示利用该特征词典对 Objective 要素的识别效果最差，其中一个重要原因在于 Objective 与 Background 在内容阐述上经常存在交叉现象，但总体识别五种要素的 F1 平均值为 0.760 6，与 2017 年王立非等在《英语学术论文摘要语步结构自动识别模型的构建》[21]一文中结合机器学习算法综合各种语言特征构建出的摘要要素识别模型的 F1 平均值 0.781 9 相差不大，证明了语义特征识别词典的识别有效性，同时利用语义特征词典对 Method 和 Result 两要素识别效果更佳。

7　结论

本研究采用了传统词典方法来识别非结构式摘要的结构要素，一方面考虑到该方法的准确性和可解释性较强，另一方面在于本研究的结果也可作为规则来完善、提升现有利用机器学习算法构建的自动标引模型效率。本研究共解决了确定结构式摘要的识别与标引方法、构建判别句子所属结构要素类别的语义特征词典和依据语义特征词典识别结构要素属性的有效性检验三个任务。语义特征词典的构建结果充分验证了最初的研究假设，说明了不同类型摘要的内在一致性，这也为未来非结构式摘要中其他要素模式的内容识别提供了一个新的思路。

本研究贡献有三：①不仅确定了结构式摘要的识别与规范化标引方法，也丰富了结构式摘要要素标签库及映射关系；②在结构要素中存在特征单词、特征动词、三词特征词块和四词特征词块等具有语义识别功能的代表性特征词汇，并构建出了包含 4 种类型词汇

chinaXiv:202304.00297v1

ChinaXiv合作期刊

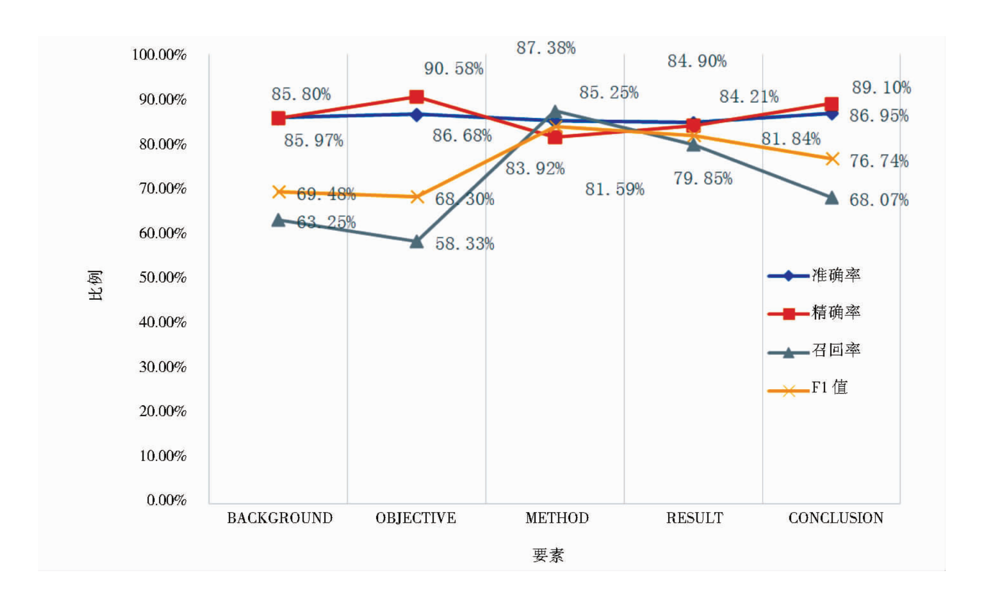


图 10 特征词典识别效果分析

chinaXiv:202304.02927v1

的语义特征词典;③语义特征词典的识别效果与当前已存在的其他摘要要素自动识别模型相当,证明了语义特征识别词典的识别有效性。

但本研究的结构式摘要样本数量有限,语义特征词典中的词汇量有待进一步扩充。同时受时间限制,未依据特征词汇的共现关系构建句型模板,需要进一步通过构建句型模板进行非结构式摘要的结构要素特征识别研究。最后在有效性测试集检测时,标注样本数量不够大,检验效果有一定的局限性。后续研究有必要进一步挖掘具有辨识功能的典型句型以及非结构式摘要文本内容深度标引的方法和智能化标引研究,为科技期刊摘要的有效利用奠定方法基础。

参考文献:

[1] ERTL N. New way of documenting scientific data from medical publications[J]. Karger gazette,1969,27(20):1 –3.

[2] 曹雁,牟爱鹏. 科技期刊英文摘要学术词汇的语步特点研究[J]. 外语学刊,2011 (3): 46 –49.

[3] DAY R A, SAKADUSKI N. Scientific English: a guide for scientists and other professionals[M]. Phoenix, AZ: Oryx, 1998:109 –125.

[4] 钱多秀, 罗媛. 基于语料库的论文摘要语步的对比研究[J]. 北京科技大学学报(社会科学版),2014,30 (2): 12 –17.

[5] GRATEZ N. Teaching EFL students to extract structural information from abstracts[M]. Belgium: ACCO,1985:123 –135.

[6] SWALES J M. Genre analysis: English in academic and research settings[D]. Cambridge: Cambridge University Press,1990.

[7] TSENG F. Analyses of move structure and verb tense of research article abstracts in applied linguistics[J]. International journal of English linguistics,2011,1 (2): 27 –39.

[8] 李涛. 科技论文的英文摘要规范化问题研究——以自然科学论文为例[J]. 辽宁工业大学学报(社会科学版),2018,20(6): 70 –73.

[9] 周志超. 中文图情期刊摘要的核心要素与逻辑结构分析[J]. 情报科学,2018,36(3):8 –12,32.

[10] DAHL T. Lexical cohesion-based text condensation: an evaluation of automatically produced summaries of research articles by comparison with author-written abstracts[D]. Bergen: University of Bergen,2000.

[11] HAYNES R B. A proposal for more informative abstracts of clinical articles[J]. Annals of internal medicine, 1987, 106(4):598 –604.

[12] NILSEN D L F, NILSEN A P. Semantic theory : a linguistic perspective[M]. Massachusetts: Newbury House Publishers,1975:1 –20.

[13] ANTHONY L E. A machine learning system for the automatic identification of text structure, and application to research article abstracts in computer science[D]. Birmingham: Birmingham University,2002.

[14] KIM S N, MARTINEZ D , CAVEDON L, et al. Automatic classification of sentences to support evidence based medicine[J]. BMC bioinformatics,2011,12(2):1 –10.

[15] FELTRIM V D, TEUFEL S. Automatic critiquing of novices' scientific writing using argumentative zoning[C]//Proceedings of AAAI spring symposium on exploring attitude and affect in text: theories and applications. 2004,3:1 –4.

[16] SILVA J, COHEUR L, MENDES A C, et al. From symbolic to sub-symbolic information in question classification[J]. Artificial intelligence review, 2011, 35(2): 137 –154.

［17］ MEENA Y K, GOPALANI D. Feature priority based sentence filtering method for extractive automatic text summarization［J］. Procedia computer science, 2015, 48(1):728－734.
［18］ GUO Y, KORHONEN A, LIAKATA I M, et al. Identifying the information structure of scientific abstracts: an investigation of three different schemes［C］//Proceedings of the 2010 workshop on biemedical natural language processings. Association for Computational Linguistics,2010:99－107.
［19］ 沈思,胡昊天,叶文豪,等. 基于全字语义的摘要结构功能自动识别研究［J］. 情报学报,2019,38(1):79－88.
［20］ U S National Library of Medcine. The NLM label list and category mappings ［EB/OL］. ［2020－01－02］. https://structuredabstracts. nlm. nih. gov/.
［21］ 王立非,刘霞. 英语学术论文摘要语步结构自动识别模型的构建［J］. 外语电化教学,2017(2):45－50,64.

作者贡献说明：

宋东桓：负责论文内容的撰写及数据处理工作；
李晨英：负责论文研究设计以及论文内容框架设计和论文写作指导及其最后审定；
刘子瑜：指导部分数据处理方法及论文的修改与完善；
韩明杰：负责部分数据处理工作及论文写作指导。

Semantic Feature Dictionary Construction of Abstract in English Scientific Journals

Song Donghuan[1,2]　Li Chenying[1]　Liu Ziyu[1]　Han Mingjie[1]
[1] China Agricultural University Library, Beijing 100193
[2] National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: ［**Purpose/significance**］ The abstract of scientific papers is a vital indexing object within information organization. Meanwhile, indexing the abstract according to certain rules is conducive for not only scientific communication or knowledge discovery, and intelligence analysis as well. Thus, how to realize auto-index accurately and quickly, for millions of unstructured abstracts existed nowadays is a crucial problem to be addressed. ［**Method/process**］ This study assumed that different categories of abstract are inherently consistent, that is, the study of structured abstract can provide a method and technical reference for unstructured abstract auto-indexing. Acting in accordance with this assumption and based on the US National Library of Medicine's structural element labeling terminology, this study accomplished mapping across abstract element classifications and proposed BOMRC system, a normalization indexing method for structured abstract. Then we collected research sample and used text mining method to analyze multiple features of structured abstract quantitatively and statistically, such as word frequency, TF-IDF value, as for dimension of words, verbs, three-word lexical chunks and four-word lexical chunks, which enabled us propose a semantic feature dictionary for structured elements. Finally, we used unstructured abstract to test the validity of the semantic feature dictionary. ［**Result/conclusion**］ The results show that the semantic feature dictionary method can effectively identify various structural elements of scientific paper abstract, and it can be used to optimize the automatic recognition model, which may be based on machine learning methods.

Keywords: scientific paper　paper abstract　structural element　semantic feature　feature dictionary

chinaXiv:202304.00297v1